processes

MDPI

Article

Multivariate Six Sigma: A Case Study in Industry 4.0

Daniel Palací-López [1,†], Joan Borràs-Ferrís [2,*,†], Larissa Thaise da Silva de Oliveria [2] and Alberto Ferrer [2]

[1] International Flavors & Fragrances Inc., IFF (Benicarló), 12580 Benicarló, Spain; daniel.palaci@iff.com

[2] Multivariate Statistical Engineering Group, Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, 46022 Valencia, Spain; ladasil@etsii.upv.es (L.T.d.S.d.O.); aferrer@eio.upv.es (A.F.)

* Correspondence: joaborfe@eio.upv.es

† These authors have equal contributions.

Received: 31 July 2020; Accepted: 7 September 2020; Published: 9 September 2020

Abstract: The complex data characteristics collected in Industry 4.0 cannot be efficiently handled by classical Six Sigma statistical toolkit based mainly in least squares techniques. This may refrain people from using Six Sigma in these contexts. The incorporation of latent variables-based multivariate statistical techniques such as principal component analysis and partial least squares into the Six Sigma statistical toolkit can help to overcome this problem yielding the Multivariate Six Sigma: a powerful process improvement methodology for Industry 4.0. A multivariate Six Sigma case study based on the batch production of one of the star products at a chemical plant is presented.

Keywords: Six Sigma; Industry 4.0; multivariate data analysis; latent variables models; PCA; PLS

1. Introduction

Six Sigma is a strategy for process improvement widely used in various sectors such as manufacturing, finance, healthcare, and so on. It is defined by Linderman et al. [1] as "an organized and systematic method for strategic process improvement and new product and service development that relies on statistical methods and the scientific method to make dramatic reductions in customer defined defect rates". Besides, Six Sigma, as a quality tool, has fostered a never-ending improvement culture based on a strong and professionalized organization for improvement, a clear and well thought methodology (DMAIC), and also powerful tools and statistical techniques to carry out the improvement projects within the DMAIC framework that has proved highly effective in a large variety of situations [2].

The DMAIC methodology in Six Sigma is a five-step improvement cycle, i.e., Define, Measure, Analyze, Improve, and Control. Reliable data and objective measurements are critical at each step of the method and, hence the statistical techniques are incorporated into the structured method as needed [1]. Traditionally, classical statistical techniques (e.g., multiple linear regression (MLR)) have been used within the DMAIC framework in a data-scarce context mainly from experimental designs. However, with the emergence of Industry 4.0 and the Big Data movement gaining momentum, data abounds now more than ever, and the speed at which they accumulate is accelerating [3]. Besides, due to the increasing availability of sensors and data acquisition systems collecting information from process units and streams, univariate or low-dimensional data matrices evolve to high-dimensional data matrices. In addition, these data often exhibit high correlation, rank deficiency, low signal-to-noise ratio, and missing values [4].

For all this, the Six Sigma statistical toolkit traditionally focused in classical statistical techniques must incorporate new approaches being able to handle complex data characteristics from this current

Industry 4.0 context. In such context, latent variable-based multivariate statistical techniques are widely recommended.

In the literature there are some examples of this integration of multivariate statistical tools into the Six Sigma toolkit. For example, Peruchi et al. [5] integrated principal component analysis (PCA) into a Six Sigma DMAIC project for assessing the measurement system, analyzing process stability and capability, as well as modeling and optimizing multivariate manufacturing processes in a hardened steel turning case involving two critical-to-quality (CTQ) characteristics. In [6], discriminant analysis and PCA were integrated into the DMAIC Six Sigma framework in order to improve the quality of oil type classification from oil spills chromatographic data.

In addition to that, latent variable regression models (LVRM) have also very attractive features, not only for their ability to build models when good predictions, process monitoring, fault detection, and diagnosis are desired (passive use), but also for being able to use this kind of Industry 4.0 data for process understanding, trouble-shooting, and optimization (active use) [4,7]. Note that for an active use causal models are required and, in contrast to machine learning (ML) and MLR models that can only yield causal models if data come from experimental designs, latent variable regression models (such as partial least squares (PLS) [8,9]) do provide causality in the reduced dimensional space of the latent variables even when using historical data corresponding to the daily production of the processes (happenstance data) [10].

This paper reinforces conclusions from previous works in the literature on how Six Sigma's DMAIC methodology can be used to achieve competitive advantages, efficient decision-making, and problem-solving capabilities within the Industry 4.0 context, by incorporating latent variable-based techniques such as PCA into the statistical toolkit leading to the Multivariate Six Sigma. An important contribution of this paper to past literature is that we advocate the use of more advanced techniques via LVRM such as PLS, and illustrate their successful integration into the DMAIC problem solving strategy of a batch production process, one of the most iconic Industry 4.0 scenarios. This type of process, although it shares many of the characteristics represented by the four V's (volume, variety, velocity, and veracity), may not really be Big Data in comparison to other sectors such as social networks, sales, marketing, and finance. However, the complexity of the questions we are trying to answer is really high, and the information that we wish to extract from them is often subtle. This info needs to be analyzed and presented in a way that is easily interpreted and that is useful to process engineers. Not only do we want to find and interpret patterns in the data and use them for predictive purposes, but we also want to extract meaningful relationships that can be used to improve and optimize a process [11] (García-Muñoz and MacGregor 2016), thus making latent variable-based techniques especially relevant, as they permit making proper use of all the data available. More specifically, this paper addresses a case study based on the batch production of one of the star products at a chemical plant.

2. Methods and Materials

2.1. Six Sigma's DMAIC Methodology

In studying Six Sigma's DMAIC methodology, De Mast and Lokkerbol [12] already commented that there are essentially two options: to study the method as it is prescribed in courses and textbooks (prescriptive accounts), or to study it as it is factually applied by practitioners in improvement Six Sigma projects (descriptive accounts). Here, this work is focused on the second option. However, it is crucial to prescribe initially the main functions of each step. Thus, a rational reconstruction of the DMAIC methodology is shown below [12]:

- Define: problem selection and benefit analysis.
- Measure: translation of the problem into a measurable form, and measurement of the current situation; refined definition of objectives.
- Analyze: identification of influence factors and causes that determine the critical to quality characteristics' (CQCs) behavior.

- Improve: design and implementation of adjustments to the process to improve the performance of the CQCs.
- Control: empirical verification of the project's results and adjustment of the process management and control system in order that improvements are sustainable.

As commented above, due to the Industry 4.0 context of the case study, we propose to incorporate latent variable-based techniques within DMAIC usual framework. To aid the reader's understanding, such techniques are described below.

2.2. Latent Variable Models

Latent variable models (LVMs) are statistical models specifically designed to analyze massive amounts of correlated data. The basic idea behind LVMs is that the number of underlying factors acting on a process is much smaller than the number of measurements taken on the system. Indeed, the factors that drive the process leave a similar signature on different measurable variables, which therefore appear correlated. By combining the measured variables, LVMs find new variables (called latent variables (LVs)) that optimally describe the variability in the data and can be useful in the identification of the driving forces acting on the system and responsible for the data variability [13].

2.2.1. Principal Component Analysis

Principal component analysis (PCA) [14–16] is a latent variable-based technique very useful to apply to a data matrix $\boldsymbol{X}$ $(N \times K)$, where N is the number of observations and K the variables measured. The goal of PCA is to extract the most important information from $\boldsymbol{X}$ by compressing the K measured variables into new A latent variables that summarize such important information. This allows simplifying the description of the data matrix and easing the analysis of the correlation structure of the observations and the variables. Thus, PCA can be used not only to reduce the dimension of the original space but also to identify patterns on data, trends, clusters, and outliers. In machine learning terminology PCA is an unsupervised method.

In order to achieve these goals PCA projects $\boldsymbol{X}$ into orthogonal directions obtaining new variables of maximum variance (i.e., principal components (PCs) also called LVs) which are obtained as linear combinations of the original variables. The decomposition carried out by PCA can be expressed as:

$$\boldsymbol{X} = \boldsymbol{T} \cdot \boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E} \tag{1}$$

where $\boldsymbol{P}$ $(K \times A)$ is the orthogonal loadings matrix, A being the number of LVs, T $(N \times A)$ is the scores matrix composed of the score vectors (columns of T), and $\boldsymbol{E}$ $(N \times K)$ is the residuals matrix. Score vectors are orthogonal to each other and explain most of the variance of $\boldsymbol{X}$. Besides, due to the orthogonality in P, the A LVs have independent contributions to the overall explained variation. To calculate the parameters in a sequential manner, the non-linear iterative partial least squares (NIPALS) algorithm can be used [17].

2.2.2. Partial Least Squares Regression

LVMs can be also used to relate data from different datasets: an input data matrix $\boldsymbol{X}$ $(N \times K)$ and an output data matrix $\boldsymbol{Y}$ $(N \times L)$, where L is the number of output variables measured. This is done by means of latent variable regression models (LVRMs), such as partial least squares (PLS) regression. Thus, LVRMs find the main driving forces acting on the input space that are more related to the output space by projecting the input ($\boldsymbol{X}$) and the output variables ($\boldsymbol{Y}$) onto a common latent space. The number of LVs corresponds to the dimension of the latent space and can be interpreted, from a physical point of view, as the number of driving forces acting on a system [18].

In contrast to classical MLR or ML techniques, PLS regression [8,9] not only model the inner relationships between the matrix of inputs $\boldsymbol{X}$ and the matrix of output variables $\boldsymbol{Y}$, but also provide a model for both. Thus, both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are assumed to be simultaneously modelled by the same LVs providing unique and causal models, which is why PLS yields causal models even with data from daily production (i.e. happenstance data not coming from an experimental design). The PLS regression model structure can be expressed as follows:

$$\boldsymbol{T} = \boldsymbol{X} \cdot \boldsymbol{W}^{*} \tag{2}$$

$$\boldsymbol{X} = \boldsymbol{T} \cdot \boldsymbol{P}^{T} + \boldsymbol{E} \tag{3}$$

$$\boldsymbol{Y} = \boldsymbol{T} \cdot \boldsymbol{Q}^{T} + \boldsymbol{F} \tag{4}$$

where the columns of the matrix $\boldsymbol{T}$ ($N \times A$) are the PLS scores vectors, consisting of the first A LVs from PLS These score vectors explain most of the covariance of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and each one of them is estimated as a linear combination of the original variables with the corresponding "weight" vector (Equation (2)). These weights vectors are the columns of the weighting matrix $\boldsymbol{W}^{*}$ ($K \times A$). Besides, the PLS scores vectors are, at the same time, good "summaries" of $\boldsymbol{X}$ according to the $\boldsymbol{X}$-loadings ($\boldsymbol{P}$) (Equation (3)) and good predictors of $\boldsymbol{Y}$ according to $\boldsymbol{Y}$-loadings ($\boldsymbol{Q}$) (Equation (4)), where $\boldsymbol{F}$ and $\boldsymbol{E}$ are residual matrices. In a predictive use, the sum of squares of $\boldsymbol{F}$ is the indicator of how good the predictive model is, and the sum of squares of $\boldsymbol{E}$ is an indicator of how well the model explains the $\boldsymbol{X}$-space.

To evaluate the model performance when projecting the n-th observation $\boldsymbol{x}_n$ onto it, the Hotelling-T^2 in the latent space, T_n^2, and the Squared Prediction Error (SPE), SPE_{x_n}, are calculated [19]:

$$\boldsymbol{\tau}_n = \mathbf{W}^{*\mathrm{T}} \cdot \boldsymbol{x}_n \tag{5}$$

$$T_n^2 = \boldsymbol{\tau}_n^{\mathrm{T}} \cdot \boldsymbol{\Lambda}^{-1} \cdot \boldsymbol{\tau}_n \tag{6}$$

$$SPE_{x_n} = (\boldsymbol{x}_n - \boldsymbol{P} \cdot \boldsymbol{\tau}_n)^{T} \cdot (\boldsymbol{x}_n - \boldsymbol{P} \cdot \boldsymbol{\tau}_n) = \boldsymbol{e}_n^{\mathrm{T}} \boldsymbol{e}_n \tag{7}$$

where $\boldsymbol{e}_n$ is the residual vector associated to the n-th observation, $\boldsymbol{\Lambda}^{-1}$ the ($A \times A$) diagonal matrix containing the inverse of the A variances of the scores associated to the LVs, and $\boldsymbol{\tau}_n$ the vector of scores corresponding to the projection of the n-th observation $\boldsymbol{x}_n$ onto the latent subspace of the PLS model. T_n^2 is the estimated squared Mahalanobis distance from the center of the latent subspace to the projection of the n-th observation onto this subspace, and the SPE_{x_n} statistic gives a measure of how close (in an Euclidean way) such observation is from the A-dimensional latent space.

PLS model can also be expressed as a function of the input variables (as in a classical regression model) by substituting Equation (2) into Equation (4):

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{W}^{*} \cdot \boldsymbol{Q}^{T} + \boldsymbol{F} = \boldsymbol{X} \cdot \boldsymbol{B} + \boldsymbol{F} \tag{8}$$

where $\boldsymbol{B}$ ($K \times L$) is the PLS regression coefficient matrix. To calculate the parameters of the model in a sequential manner, NIPALS algorithm can be used [20]. In both PCA and PLS regression, NIPALS algorithm has two main advantages: it easily handles missing data and calculates the LVs sequentially (an important property from a practical point of view).

Although PLS was not inherently designed for classification or discrimination, it can be used for both purposes in the form of PLS discriminant analysis (PLS-DA) [21]. Thus, by means of PLS-DA one can explain differences between overall class properties and classify new observations. In machine learning terminology PLS and PLS-DA are supervised methods.

2.3. LVMs in Batch Processes

Batch processes operate for a finite period with a defined starting and ending point, and a time-varying behavior over the operating period. Thus, the data available on batch processes fall into three categories: summary variables ($\boldsymbol{X}^{SV}$) characteristics of each batch such as initial conditions, charge of ingredients, shift, operator, or features from the trajectories of the process variables throughout batch evolution; time-varying process variables throughout the batch evolution (trajectory variables—$\boldsymbol{X}^{TV}$; and the CQCs of final product ($\boldsymbol{Y}$)). The nature of these data is represented in Figure 1.

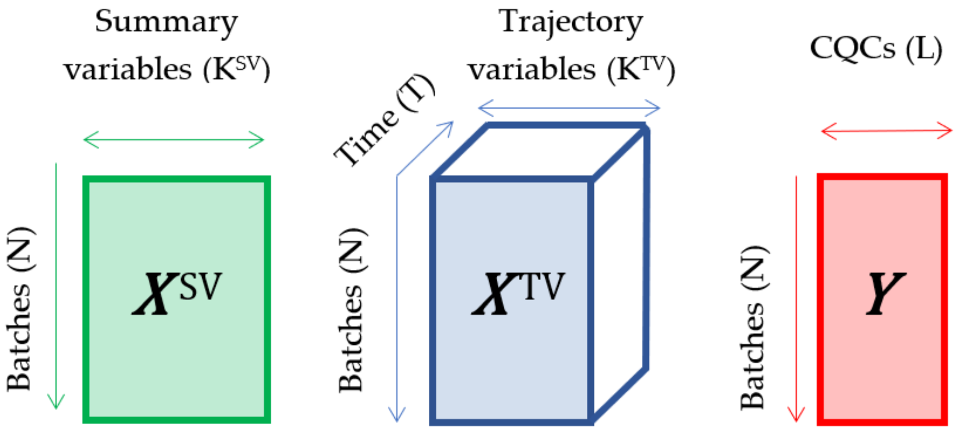

Figure 1. Nature of the batch data set.

The presence of very large arrays, missing data, and the fact that there is temporal and contemporaneous correlation among all variables, provide a strong motivation for using latent variable-based techniques in such data. However, these techniques are based on bilinear modeling approaches and, therefore, a priori they are not able to handle that three-dimensional (3D) $\boldsymbol{X}^{TV}$ arrays. For that, several approaches for analyzing batch data have been proposed where the differences mainly revolve around how to treat $\boldsymbol{X}^{TV}$ [22]. For example, the simplest approach is to extract meaningful landmark features from the trajectories and recording the values of such features for each batch run in a 2D data matrix.

More sophisticated approaches are based on unfolding the 3D $\boldsymbol{X}^{TV}$ in a 2D matrix. One approach is the batchwise unfolding (BWU) [23], that is, to unfold $\boldsymbol{X}^{TV}$ such that all the information for each batch is contained in one row. The data are then mean centered and scaled for each column. Mean centering removes the mean trajectories of each variable, eliminating the main nonlinearity due to the dynamic behavior of the process, and focusing on the variation of all the variables about their mean trajectories. This approach allows exploiting the complex auto and lagged cross correlation structure (i.e., dynamics) of the batch process. Performing any subsequent PCA or PLS analysis on this matrix then summarizes the major sources of variation among the different batches and allows efficient batch-to-batch comparison. This approach also allows for incorporating summary variables ($\boldsymbol{X}^{SV}$) and final product quality variables ($\boldsymbol{Y}$) associated with each batch when performing either a PCA or a PLS analysis [19].

Another approach is to unfold the data observation-wise (OWU) [24] with each row corresponding to an observation at some time in each batch, and each column corresponding to the variables measured. Mean centering by column (variable) then simply centers the origin of each variable about zero, but does not remove the average trajectories. The variation remaining is the total trajectory variation for each variable. Performing PCA or PLS (using local batch time or a maturity variable as $\mathbf{y}$) on these OWU data finds a smaller number of components that summarize the major behavior of the complete trajectories of the original variables. It does not initially focus directly on the differences among batches as BWU does, but on summarizing the variable trajectories [22]. The latter can be overcome by carrying out a second model, being similar to the single model BWU, except that it is based on the unfolded score matrix $\boldsymbol{T}$ of the OWU (i.e., OWU-TBWU). The main motivation of the OWU-TBWU approach is dimensionality reduction before BWU stage, which is required when the number of trajectory variables is very large (e.g., when online analytical sensors such as Mass or NIR spectrometers are used). However, the modeling of the OWU data by a single PLS model works well if there is no important dynamics in the process and the instantaneous correlation structure (i.e., the correlation structure or the variables at the same time) remains stable throughout the whole batch evolution. Nevertheless, such assumption does not seem to be realistic at the chemical batch processes analyzed in this case study and, hence, BWU approach is used in this work.

Moreover, batch trajectories are dependent on time (more specifically on the pace the batch is run) and they are rarely synchronized (i.e., the key process events do not overlap at the same time in all batches) [25]. To compare these batch histories and apply statistical analysis one needs to reconcile the timing differences and the key process events among these trajectories [26]. This can be achieved using the dynamic time warping (DTW) method with only a minimal amount of process knowledge [27]. This method nonlinearly warps all batch trajectories to match as closely as possible that of a reference batch. Figure 2 illustrates the BWU approach followed by DTW synchronization.

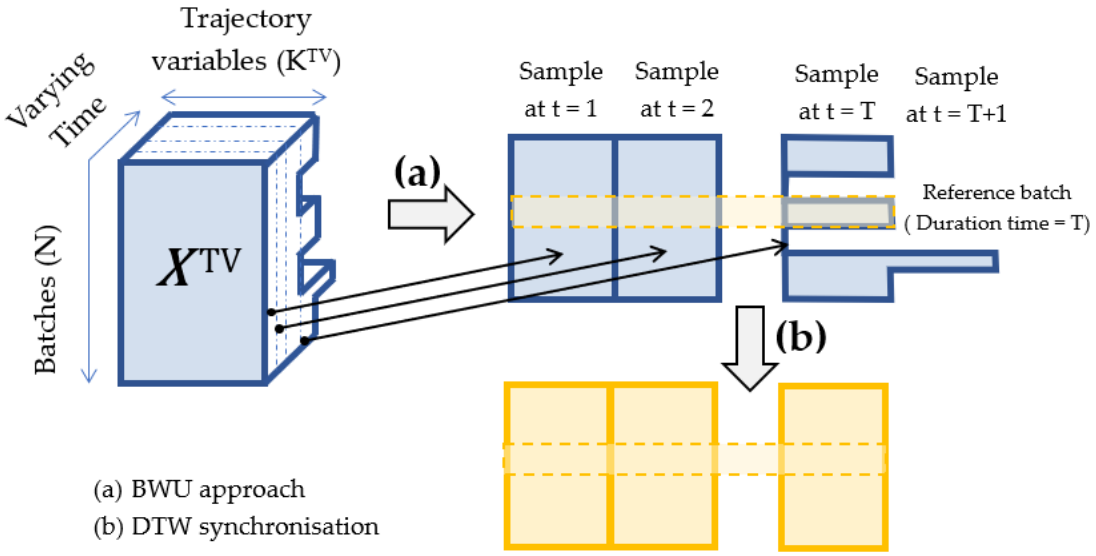

Figure 2. (**a**) Batchwise unfolding (BWU) approach and (**b**) dynamic time warping (DTW) synchronization of the 3D X^{TV}.

Note that, once X^{TV} is unfolded and synchronized as seen in Figure 2, it is concatenated with X^{SV}, resulting in the matrix X to be used in the PCA or PLS model.

2.4. Software

The software packages used are as follows:

- Multivariate Exploratory Data Analysis (MEDA) Toolbox (for Matlab) [28] for variable and batch screening, and imputation of missing data within a batch.
- MVBatch Toolbox (for Matlab) [29] for batch synchronization.
- Aspen ProMV for calibration by using synchronized batch data, and data analysis.
- Minitab for control chart plotting.

3. Results

In this section, the results from applying each of the DMAIC steps (Define, Measure, Analyze, Improve, and Control) to the process are shown, each in their own subsection. Note that latent variable-based methods such as PCA or PLS are used, instead of more classical ones such as MLR or even ML. Therefore, the tools implemented in some of the steps of the DMAIC cycle differ from more traditional approaches, but the original purpose of each stage remains.

3.1. Define

The aim of this stage is to identify opportunities for improvement that lead to e.g., an increase in benefits, reduced costs or losses, a mitigation of the environmental impact, etc. This requires pinpointing observed problems, framing them within the context of the corresponding processes, evaluating the costs and benefits of addressing them, and locating the most appropriate people to do it given the existing constraints on time and resources.

In this Six Sigma project, the focus was set on the purity (before separation) and volume of production (after purification) of one of the star products at the chemical plant where the continuous improvement program was implemented. This came as a result of an observed increase in the variability of this product's purity, due mostly to significantly lower values (compared to previous operations) being obtained once every four batches, approximately. This also meant a decrease in its average value of around 1%, starting on September of 2014, with an estimated monetary loss of more than 100,000 €/year with respect to previous years.

Figure 3a illustrates the 'Suppliers, Inputs, Process, Outputs, Customers' (SIPOC) diagram identifying the supplier (reaction 1) and inputs (one of the outputs from reaction 1) for the specific process under study (reaction 2), as well as its outputs (with primary focus on the composition of the so-called subproduct 2) and the customer (reaction 3). Figure 3b corresponds to a simplified process block diagram for reaction 2, in which four numbers have been included to indicate the points of the process where the process variables and critical to quality characteristics (CQC) relevant to this work are routinely measured.

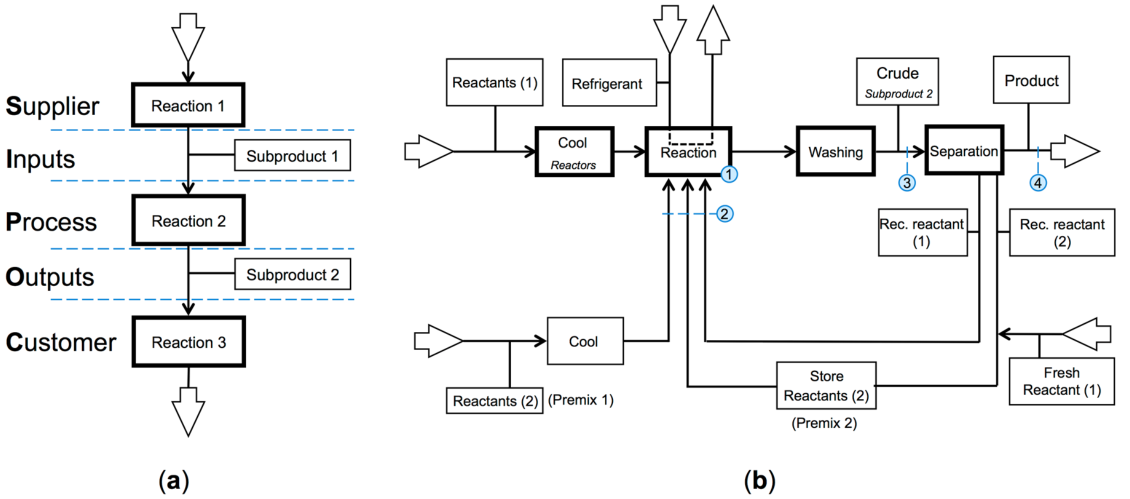

Figure 3. (**a**) 'Suppliers, Inputs, Process, Outputs, Customers' (SIPOC) diagram; (**b**) simplified process block diagram for reaction 2.

Due to confidentiality reasons, the name of the reactants, products, and subproducts cannot be disclosed. However, it must be noted that both 'Reactants (1)' and 'Fresh Reactant (1)' come from reaction 1. Likewise, Premix 1 and Premix 2 are blends constituted by the same chemical species, these being freshly produced raw materials for Premix 1, and one of them (Rec. reactant (2)) having been recovered after the washing and separation following reaction 2. For each batch, each reactor was fed Premix 1 or Premix 2, but never a blend of both. A quick look at already available data revealed that the main concern that motivated this Six Sigma project was related to a loss of purity (in terms of the desired chemical species) of the Crude, (i.e., unrefined product) and therefore a lower amount of Product arriving to reaction 3 per batch of reaction 2.

The project team was constituted by the authors of this paper (Six Sigma Black Belts), championed by two high-profile members from the company and supported by three experts in the process from the technical team. Regarding the planning, the project was estimated to require over six months to be completed. Experimenting in a laboratory or at pilot-plant level was not recommended, since no proper scaling could be done. Additionally, altering or interrupting the production of this process was not allowed to any extent, and therefore experimenting on the plant itself was not an option either. Due to this, only historical data from past production could be used.

3.2. Measure

During this stage, as much available data as possible was collected, and their validity assessed. With these data, an evaluation of the initial situation was done, and potential causes were looked at for the issue that motivated the project.

3.2.1. Available Data

Data from a total of 17,147 batches produced in two different reactors during a nine years period was available, containing information about:

- the averaged values for three process variables (x_1 to x_3) for each batch, measured at point (1) in Figure 3 (i.e., the corresponding reactor);
- amounts (x_4 to x_7) and proportions (x_8 to x_{11}) of some of the most relevant reactants involved in the reaction, measured at point (2) in Figure 3 (i.e., before being introduced into the reactor);
- four categorical variables indicating whether a batch was produced in the first or second reactor (x_{12}) and the use or not of an auxiliary piece of equipment (x_{13}), registered at point (1) in Figure 3; and whether an excess of accumulated reactant had been recovered or not (x_{14}), and whether Premix 1 or Premix 2 had been fed to the reactor for the corresponding batch (x_{15}), registered at point (2) in Figure 3;
- the evolution along the complete duration of each batch for 11 process variables (x_{16} to x_{26}), measured at point (1) in Figure 3, and;
- information on 10 CQC (y_1 to y_{10}), including the purity of the product of interest (y_8), measured at point (4) in Figure 3; and the measure of the total amount of crude coming out of the process (y_4), and its estimation through mass balance (y_6), measured/registered at point (3) in Figure 3.

The first three groups of variables will be referred to as ‘summary variables’, since they provide summarized information of each batch (e.g., average observed values or setpoint values), disregarding their variation during the evolution of the chemical reaction until completion and discharge of the reactor. The fourth group, on the contrary, is comprises ‘trajectory variables’ that may show differences in the evolution of the corresponding process conditions among batches even when their average or target values coincide.

Although further experimentation may be suggested during this step to enrich the database used in following stages, such experimentation is not possible in this case, as previously stated, and therefore no such approach will be addressed in this section.

3.2.2. Validation of the Data

In order to detect potential outliers, a PCA model with two latent variables (LVs) was fitted using all available data for the ‘summary variables’ as provided, resulting in a model that explained 17% of the variability of these data. Adding any more LVs provides no additional information that is useful at this stage, and instead results in PCA models with lower explanatory and predictive capabilities, and less ability for the detection of outliers. A representation of the SPE of all observations in the database resulted in Figure 4.

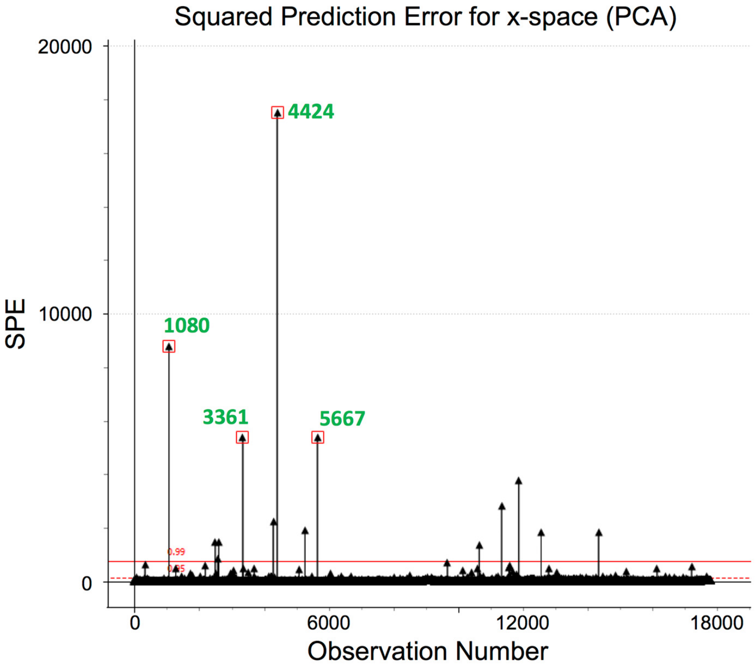

Figure 4. SPE for all observations in the dataset with the summary variables and critical to quality characteristics (CQCs) for a principal component analysis (PCA) model fitted with two LVs [$R^2(X)$ = 17%], SPE 95% (dotted red line) and 99% (continuous red line) confidence limits, and the four observations with highest SPE values highlighted.

This plot allows quickly detecting observations that do not abide by the correlation structure found by the PCA model in the dataset for the 'summary variables'. A contribution plot, such as the one in Figure 5, provides additional information regarding which variables are responsible for the high SPE value for the corresponding observation. In this case, variable x_4 presents an abnormally high value for observation 1080, not following the correlation structure found in the data by the PCA model.

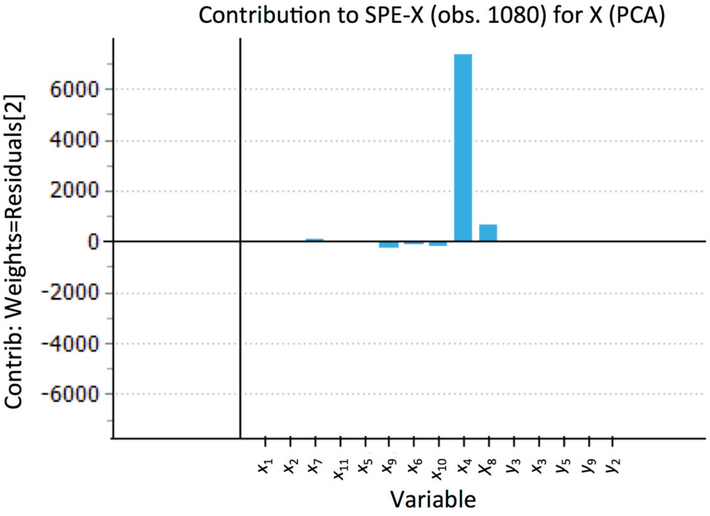

Figure 5. Contribution plot for observation 1080, seen in Figure 4 to have a SPE value significantly above the 99% confidence limit for a PCA model fitted with two LVs [$R^2(X)$ = 17%]. Variable x_4 is seen to be the biggest contributor to the SPE value.

An in-depth analysis of the factors that contribute to the high SPE values of all observed outliers allowed curating the database to either correct wrongly registered data or to eliminate outliers before continuing with the analysis. Consequently, the dataset was reduced from 17,147 original batches to 16,813.

The same procedure was followed for the dataset containing the 'trajectory variables', but no outliers were found among these data other than the ones already identified with the 'summary variables'. As a consequence, these observations were discarded before continuing.

On the other hand, process variable x_5 was found to present almost no variability in the dataset and was also discarded before going on. Additionally, variables 'day', 'month', and 'year' were included only as labels with which observations could be colored, in order to identify possible patterns, stationary effects or changes with time without artificially biasing the model to account for these variables. However, no clustering or displacement of the observations was detected this way, and the presence of outliers was not found to be correlated to these time-related variables.

3.2.3. Quantified Initial Situation and Potential Causes of the Observed Problem

Once outliers were eliminated from the dataset, the starting point of the project was determined according to the remaining information. Figure 6 shows the evolution of the purity of the product of interest (y_8) with time for both reactors. The superimposed dashed blue lines mark the separation between batches produced before and after September 2014.

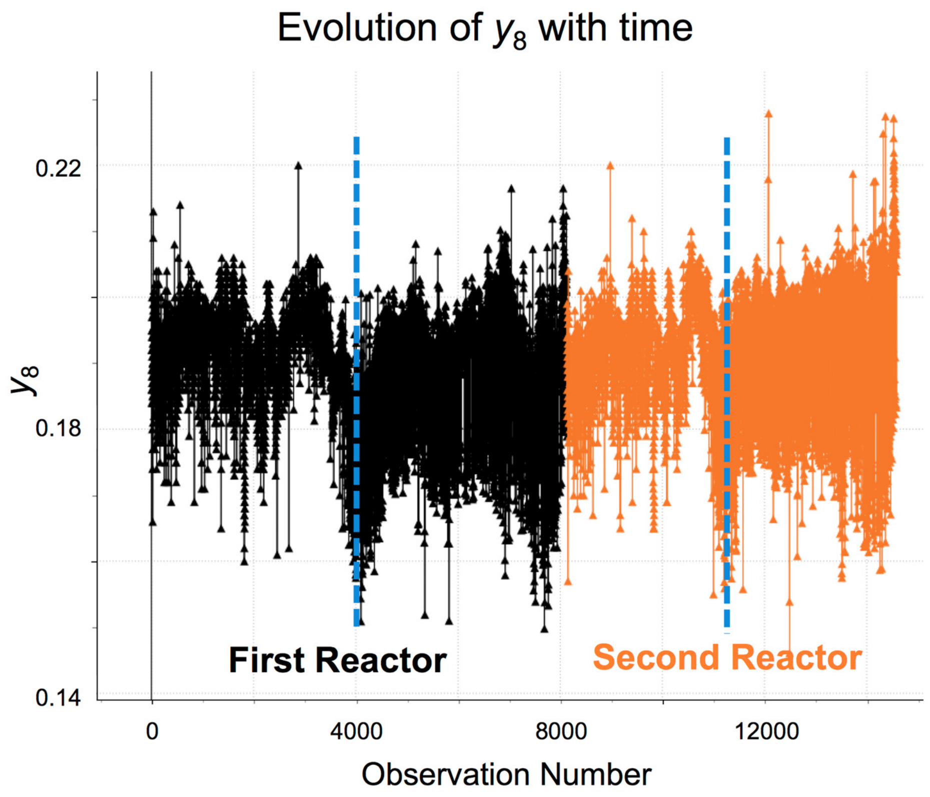

Figure 6. Evolution of the purity of the product of interest (y_8) with time for the first (black) and second (orange) reactors; the dashed blue lines separate batches produced before (left side) and after (right side) September 2014.

The average value for y_8 after September 2014, compared to before, was around 0.1% lower, while its standard deviation had increased to 1.008% (1.47 times that of past batches). Both changes (in average value and variability) were found to be statistically significant (p-value < 0.05), which corroborated, at least partially, the concerns expressed by the technicians at the start of the project. When asked, they mentioned that several changes had taken place at some point during 2014, such as the addition of an auxiliary refrigerating system to one of the reactors, the way the reactants were fed or the recovery of some amount of unreacted raw materials after each reaction.

3.3. Analyze

The main goal of this stage was to identify which process parameters have a significant effect on the product's purity, evaluate the nature of their effect (antagonistic or synergic), and how they relate to each other. In order to achieve it, a PCA model was fitted, and presented in Section 3.3.1, with all summary variables and CQC to explore the correlation structure among them in the database, and to detect clusters of batches that operated in similar way in the past. Afterwards, a PLS-regression model permits predicting the CQCs from the summary variables, and was used to determine which of these factors have a significant effect on the purity of the product of interest (y_8), as seen in Section 3.3.2. Once these variables are identified, a PLS-DA was performed considering the trajectory variables, to assess which of them are responsible for the observed differences between batches with higher and lower performance, as illustrated in Section 3.3.3.

3.3.1. Principal Component Analysis of the Summary Variables and CQCs

This first exploratory analysis was aimed at providing relevant information regarding the existing correlation structure among summary variables and CQCs, and detecting clusters of batches operating in similar conditions and/or providing similar results. Given that outliers were already eliminated from the dataset, a PCA model with five LVs [$R^2(X) = 76\%$] could be directly fit. Additional LVs beyond the fifth corresponded to either variation of individual variables independently of others, or variations not related to the CQCs, and were therefore not considered relevant to the goal of the project.

The Hotelling-T^2 values for the observations used to fit this model can be seen in Figure 7a. Here, batches were colored by variable x_{13} (black: 0; orange: 1). In Figure 7b, the scores plot of LV2 (explaining 20% of the variability of the data) versus LV1 (explaining 22% of the variability) is shown, such that the left red cluster corresponds to the observations in orange in Figure 7a, and the rest correspond to observations in black in Figure 7a. Figure 7c, where the loadings for the variables in the two first latent variables are represented, allows the interpretation of this clustering. In it, variables x_6 and x_{10}, and $x_{13} = 1$, can be seen on the left side, with values close to zero in the second component, while variables x_4 and x_8, and $x_{13} = 0$, are found in the opposite side. This provides two valuable pieces of information (which Figure 7d illustrates, too):

- Variables x_4 and x_8 present higher values for the batches in the blue cluster in Figure 7b, and lower values for the batches in both red clusters, while the opposite is true for variables x_6 and x_{10}. Variable x_{13} takes the value 0 for batches in the blue cluster and in the rightmost red cluster, and 1 for the leftmost red cluster.
- Variables x_4 and x_8 are positively correlated, as are variables x_6 and x_{10}, and the former couple is negatively correlated with the later.

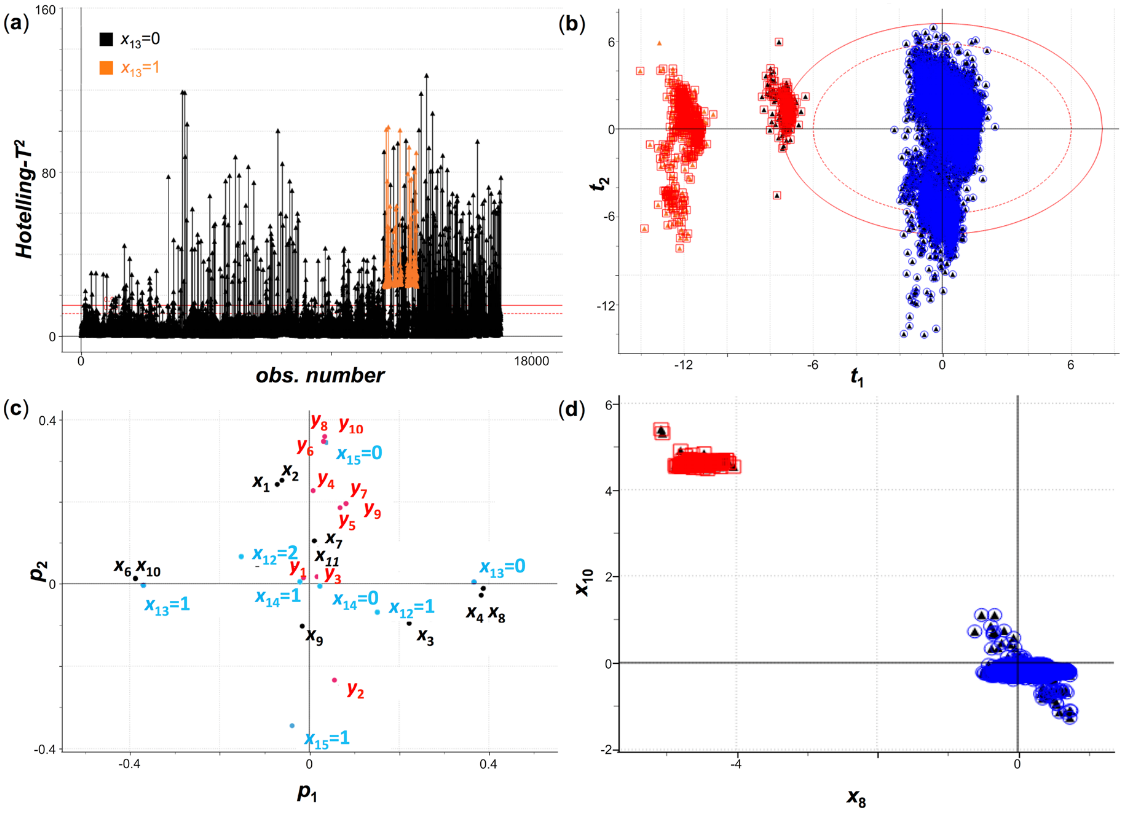

Figure 7. (**a**) Hotelling-T^2 plot of the observations in the dataset with the summary variables and CQCs for a PCA model fitted with five LVs [$R^2(X) = 76\%$], T^2 95% (dotted red line) and 99% (continuous red line) confidence limits, colored by x_{13} (black: 0; orange: 1); (**b**) scores plot for the two first LVs (t_2 vs. t_1) showing three clusters of observations: red circled orange dots associated to $x_{13} = 1$, above average values for x_6 and x_{10} and below average values for x_4 and x_8; red circled black dots associated to $x_{13} = 0$, above average values for x_6 and x_{10} and below average values for x_4 and x_8, and; blue circled black dots associated to $x_{13} = 0$, below average values for x_6 and x_{10} and above average values for x_4 and x_8; (**c**) loadings plot for the two first LVs (p_2 vs. p_1) with CQCs in red, continuous process variables in black, and binary process variables in cyan; (**d**) scatterplot for x_{10} vs. x_8, using the same color code as in Figure 7b.

Note that, more generally, the relationships among all process variables and CQCs in the dataset used to fit the PCA model can also be assessed by looking at the loading plots. In this plot, if the corresponding LVs explain a relevant percentage of the model variability, variables lying close to each other (and far away from the center) will tend to show positive correlation; while if they lay at the opposite site in the plot they will tend to show negative correlation. Figure 7d can be resorted to for carrying out such analysis (latent variables three to five do not, in this case, alter this interpretation). This way, in addition to the aforementioned correlations, positive correlations were found between variables y_4, y_6, y_8, and y_{10}, and variables x_1, x_2, and $x_{15} = 0$, as well as between x_3 and variables y_5, y_7, and y_9. On the other hand, negative correlations were found between y_2, and all other CQCs except for y_1 and y_3, as well as between x_3 and variables x_1 and x_2, and between x_9 and variables x_7 and x_{11}. More importantly, no clear correlation was found between y_8 and variables x_8 to x_{11}. Bivariate dispersion plots for each pair of variables were used to visualize each of these relationships (or lack thereof), and also allowed detecting that not only was $x_{15} = 0$ positively correlated with y_8, but that the intensity of the positive/negative correlations between y_8 and other process variables and CQCs varied when $x_{15} = 1$ (Premix 2 fed to the reactor) with respect to $x_{15} = 0$ (Premix 1 fed to the reactor). As an example, the positive correlation between x_2 and y_8, as well as the relationship between x_{15} and y_8, and the interaction effect between x_2 and x_{15}, are shown in Figure 8.

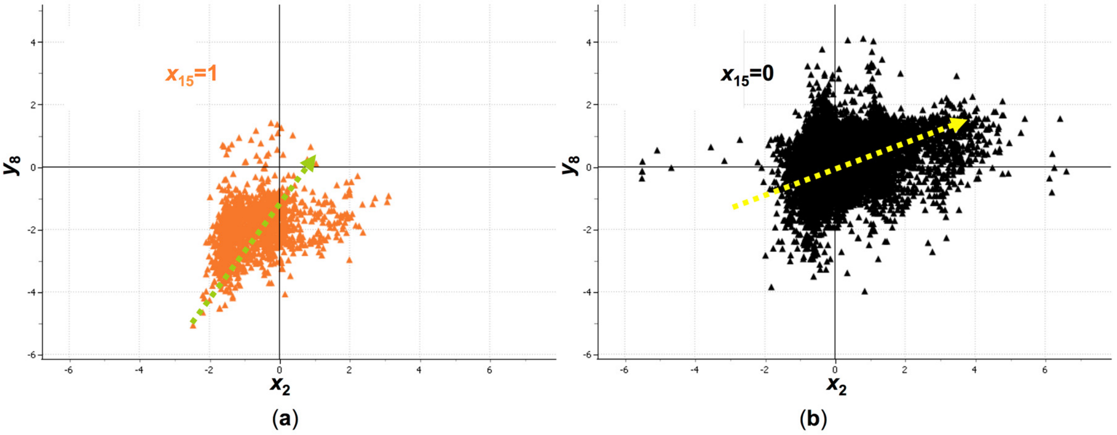

Figure 8. Scatter plot for y_8 vs. x_2, for (**a**) x_{15} = 1, and the approximate direction of maximum variability indicated by a green arrow, and (**b**) x_{15} = 0, and the direction of maximum variability indicated by a yellow arrow.

From Figure 8a,b, the positive correlation between x_2 and y_8 can be immediately confirmed. Furthermore, the cluster of batches for which x_{15} = 0 presents higher values (on average) than those for which x_{15} = 1. This is coherent with the conclusions extracted from Figure 7c Additionally, however, the slopes of the green arrow in Figure 8a and the yellow one in Figure 8b differ, pointing to a stronger, more positive correlation between x_2 and y_8 when x_{15} = 1, compared to their weaker, but still positive, relationship when x_{15} = 0. Therefore, it can be suspected that an interaction exists between x_2 and x_{15}.

3.3.2. Partial Least Squares Regression to Predict the CQCs from the Summary Variables

This analysis was performed in order to identify the sources of variability of the process most related to the product's purity (i.e., variables y_4, y_6, and y_8). This required confirming previous results and quantifying the relationship between the summary variables and the CQCs. For the sake of brevity, only the results regarding the established predictive model for y_8 will be shown in this section, as those to predict y_4 and y_6 provide the same overall conclusions. The potential effects of time related variables ('month' and 'year') and interaction effects between the categorical variables x_{12} to x_{15} and other summary variables were also considered initially. However, no statistically significant differences in the CQCs were found between reactors, and the effect of variables 'month' and 'year' was not statistically significant either. Furthermore, all interaction effects were discarded for the same reason, except for the interactions of both levels of x_{15} with variables x_1 and x_2. Figure 9 presents the coefficients of the resulting PLS-regression model fitted with two LVs [$R^2(Y)$ = 25.86%; $Q^2(Y)$ = 25.81%] to predict y_8.

One important consideration in this analysis is that variable x_3, which is known to be critical in the process, did not appear to have a significant effect on the most relevant CQCs. This is partially because this is a very strictly controlled variable. On the other hand, variables x_1 and x_2, with which x_3 is negatively correlated, are found to apparently have a significant positive effect on y_8. From this, and according to the technicians' knowledge of the process, one may conclude that it is x_3 that has a statistically significant, negative effect on y_8, but one should be cautious given that x_1, x_2, and x_3 do not vary independently. This is illustrated in Figure 10, where batches with higher values for x_1 also present higher values of y_8, on average (i.e., for lower values of x_1, batches with similar and smaller values of y_8 are observed), which points to the positive correlation between x_1 and y_8. On the other hand, batches with the highest values of x_1 only operated at values of x_3 close to its historical minimum, which also illustrates the negative correlation between x_1 and x_3. This could also be seen in the loadings plot of the PCA in Figure 7c. To disentangle this potential aliasing some experimentation should be run in the future, when/if possible.

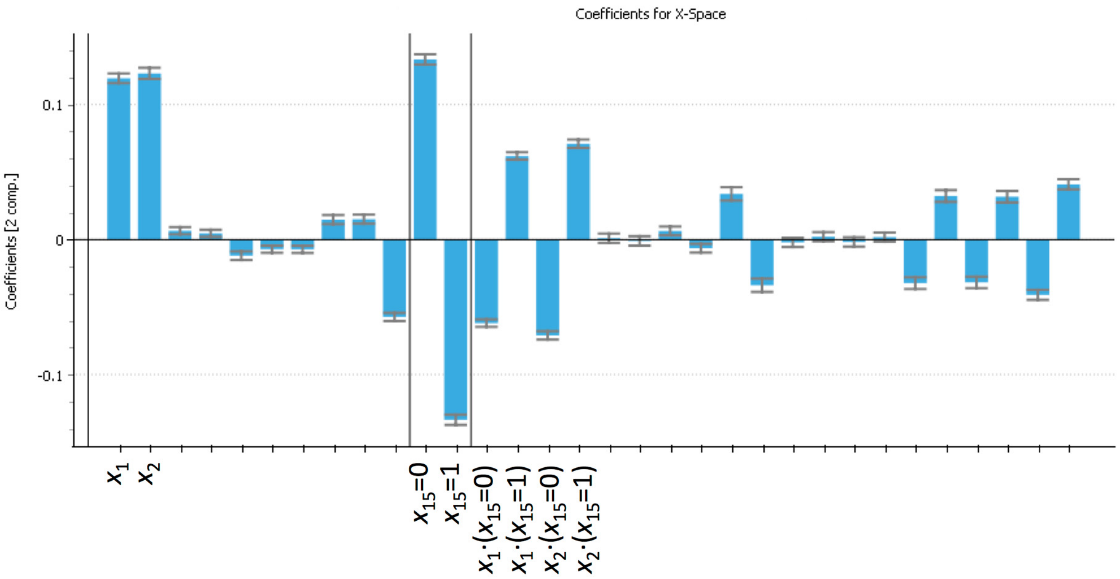

Figure 9. Regression coefficients of the partial least squares (PLS)-regression model fitted with two LVs [$R^2(Y) = 25.86\%$; $Q^2(Y) = 25.81\%$] to predict y_8 from the summary variables and their interactions with x_{15}.

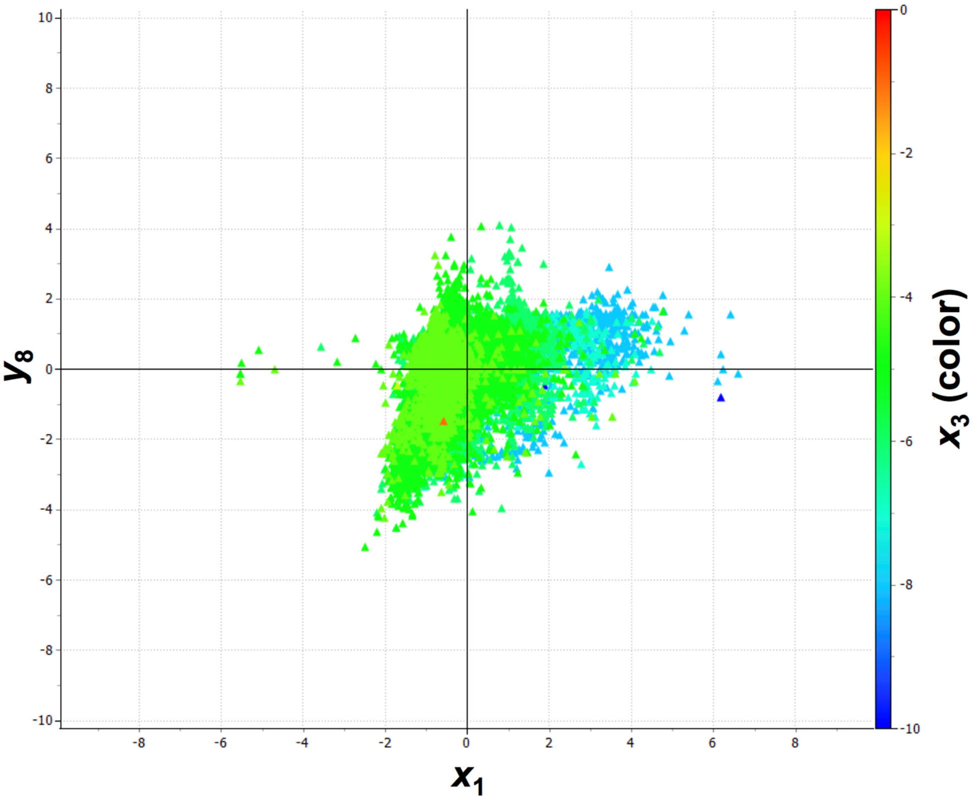

Figure 10. Scatter plot for y_8 vs. x_1, with the observations colored according x_3.

Nevertheless, the negative relationship between $x_{15} = 1$ (using Premix 2) and y_8, already seen in Figure 8, was confirmed once more, as seen in Figure 9, and so finding what is being done differently in such case becomes relevant.

3.3.3. PLS-Discriminant Analysis to Identify Differences in Batches Using Premix 1 and Premix 2

Since x_{15} seems to be one of most important variables affecting the purity of the product, y_8, conclusions obtained by previous analysis were confirmed by means of a PLS-DA model, which was resorted to for finding which variables are responsible for the differences in how batches operated when Premix 2 ($x_{15} = 1$) was fed into the reactor, compared to when Premix 1 ($x_{15} = 0$) was used. This analysis was carried out considering both the summary and trajectory variables, but only the results with the trajectory variables (x_{16} to x_{26}) are illustrated here, for the sake of both brevity and clarity. Figure 11 shows the separation between both clusters of batches in the latent space, while Figure 12 presents the model coefficients associated to each process variable, included the 'warping profile' that results from aligning the trajectories, for a PLS-DA regression model to predict $x_{15} = 1$. This model was fitted with eight LVs, as this number of LVs provided the model with the most discriminant power [$R^2(Y) = 79.80\%$; $Q^2(Y) = 71.90$].

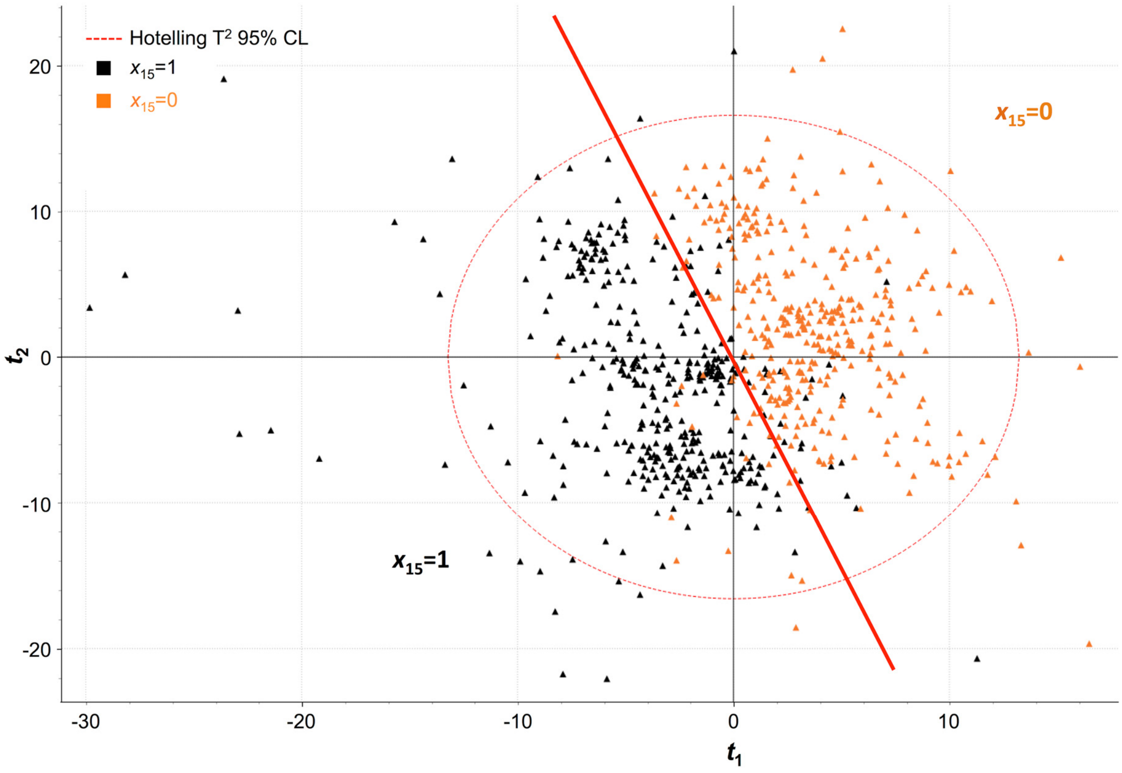

Figure 11. Scores plot for the two first LVs (t_2 vs. t_1) for the PLS-discriminant analysis (DA) regression model fitted with eight LVs [$R^2(Y) = 79.80\%$; $Q^2(Y) = 71.90\%$], showing the separation between batches with $x_{15} = 0$ (orange; right side of the red straight line) and $x_{15} = 1$ (black; left side of the red straight line).

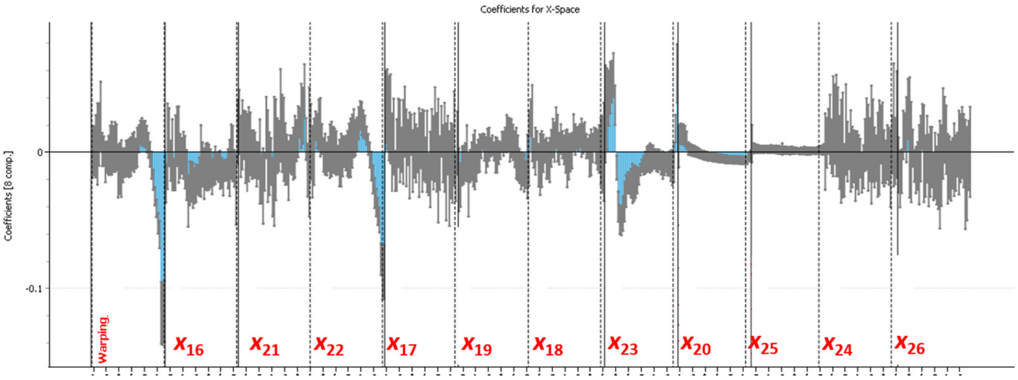

Figure 12. Model coefficients (blue bars) and the confidence interval for a 95% confidence level (grey intervals) associated to each variable for the PLS-DA regression model to predict x_{15} = 1 from variables 'warping profile' and x_{16} to x_{26}, fitted with eight LVs [$R^2(Y)$ = 79.80%; $Q^2(Y)$ = 71.90]. Positive values indicate that higher values for the corresponding variable at that point in the batch are expected, on average, for batches with x_{15} = 1 (Premix 2 used).

From Figure 12 it is concluded that batches where Premix 2 was fed to the reactor (x_{15} = 1):

- Proceeded, at the latest stages of the batch, faster than those where Premix 1 was fed instead (x_{15} = 0), as seen by the negative values for variable 'warping' near the end of the batch duration.
- Presented lower (and decreasing) values for variable x_{22} (ingredient flow) at the end of the batch.
- Operated at higher values for variable x_{23} (ingredient temperature) at the start of the batch, but lower during the middle part of the batch.

It is worth noting that, although the technicians at the plant were not surprised by the discrepancy in the values observed for variable x_{23} at the start of the batch for the two clusters, the values during the middle part was, according to them, opposite to their expectations.

3.4. Improve

As a result of the analyses performed and summarized in the previous section, the team responsible for this process was able to pinpoint a specific behavior in the process potentially related to the loss of purity/production volume of the desired product (y_4, y_6, and y_8) when Premix 2 was fed into the reactor. When compared to one another, the average value for y_8 for x_{15} = 1 was found to be 1.9332% lower than for x_{15} = 0, and the standard deviation for y_8 for x_{15} = 1 was also 1.12 times that for x_{15} = 0, with both differences being statistically significant (p-values < 0.05). A preliminary confirmation of this was provided by the historical data itself, according to which Premix 2 was fed to the reactor for the first time on September 2014, matching the increase in variability and drop in average value for y_8. While the technicians at the plant knew about this operational change (which they themselves implemented), they initially argued that such modification did not justify the change in the process outputs, but the unexpected behavior of variable x_{23} (see Figure 11) pointed to either a contamination issue with Premix 2, or to other, not registered process variables as potential causes. Consequently, they decided to further investigate them and, once detected, to correct the actual issue (the details on how this was done is not disclosed for confidentiality issues) and standardize the treatment of both Premix 1 and Premix 2 to make them equal. The solution was validated during the next two months, and remains a success to this day, with estimated benefits/savings above 140,000 €/year (40,000€ higher that the initial estimation in the 'Define' step).

3.5. Control

Once the causes of the loss in productivity were detected and addressed, an instrumental monitoring scheme was implemented in the plant to detect possible deviations in the process variables, and specially with regards to Premix 1 and Premix 2. This monitoring system was designed in a way that avoids expensive and time-consuming tests (details not disclosed to the project team), the lack of which allowed the actual cause of the issue to go unnoticed before the Six Sigma project was carried out. Currently, the use of multivariate statistical analysis techniques for monitoring purposes supports this instrumental monitoring scheme by quickly signaling out-of-specification batches and delimiting the range of potential causes for such events.

As an example of the achieved improvement, as well as an appropriate Individual-Moving Range (I-MR) control chart to be used for monitoring y_8 after the project goal was achieved, Figure 13 illustrates the evolution of variable y_8 before and after improvement, for batches where Premix 2 was fed to the reactor.

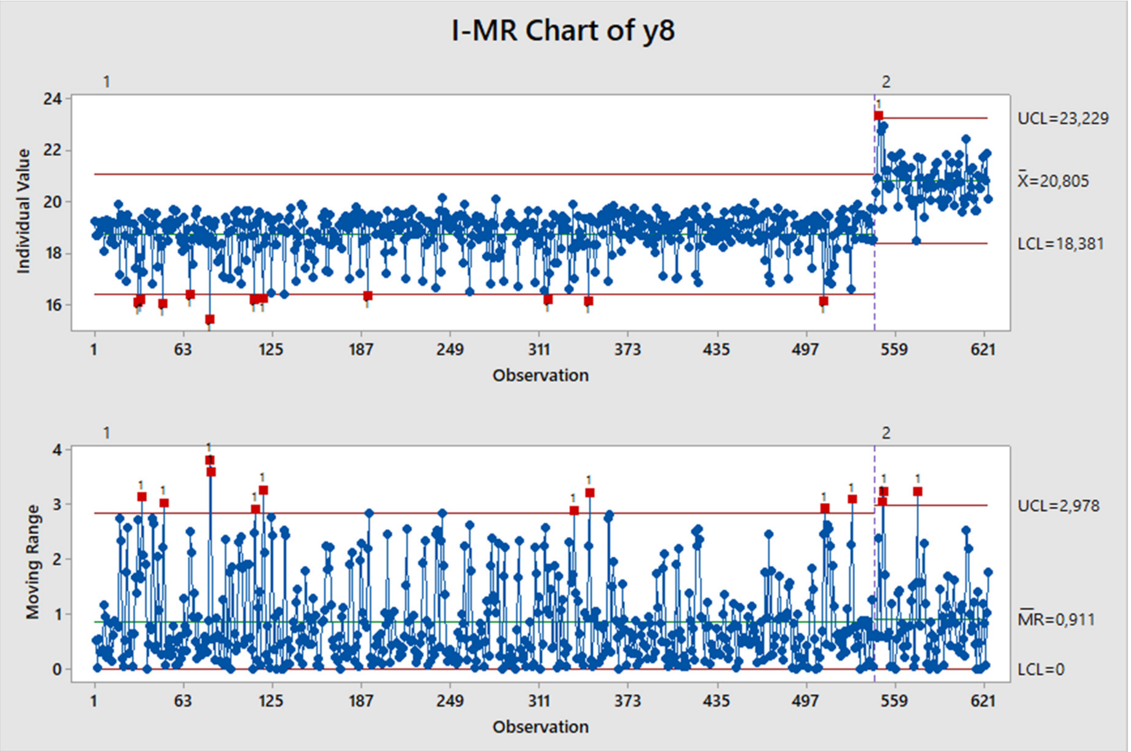

Figure 13. Individual-Moving Range (I-MR) control chart for variable y_8, for batches where Premix 2 was fed into the reactor, before (1) and after (2) improvement in accordance to the conclusions of the Six Sigma project.

4. Discussion

While the success of the Six Sigma methodology has already been documented in the past in numerous industrial case studies, the tools used (mainly based on linear regression and simple graphical displays) are usually those suitable for scenarios where not much information is available yet, a relatively limited number of factors are involved or relevant, and/or experimentation can be carried out to a minimum yet significant extent. In the problem addressed in this project, related to a batch production process, due to the nature of the data registered, typical of Industry 4.0, none of these apply and, therefore, alternative methods had to be resorted to.

In particular, latent variable-based methods such as PCA, PLS, and PLS-DA, applied to historical (i.e., not from DOE) data of both summary variables and trajectory variables (usually just referred to as batch data) were able to extract valuable information to pinpoint the actual causes of the loss of productivity in a real case study. These tools were also implemented for troubleshooting purposes in the future.

In contrast with these latent variable-based methods, data from DOE would have been required to use tools such as linear regression or machine learning tools to infer causality, which is needed for process understanding and optimization purposes. However, as it is typical in Industry 4.0 no data from DOE were available. As an example, when linear regression was applied to the available historical data, due to the highly correlated regressors (process variables), different models using different regressors and having different weights or coefficients on them gave nearly identical predictions and similar to PLS model, but failed to properly identify the relationships between x_1, x_2, x_3, and y_8, as well as the existing interaction effect between the use of Premix 1 (x_{15} = 0)or Premix 2 (x_{15} = 1) and other process variables, such as the ones shown in Figure 8. Had any of such linear regression models been used in the 'Improve' step of the DMAIC for process improvement, a different set of process operating conditions would have been advised with a high probability of not being feasible in practice, as a result of e.g., the actual relationships between x_1, x_2, and x_3 going unnoticed. Furthermore, a lesser degree of improvement would have been achieved, presumably, if the interaction between e.g., x_2 and x_{15} had not been discovered. Therefore, this constitutes a clear example of the dangers of resorting to more basic linear regression (and also machine learning) techniques for process optimization in scenarios they are not suitable for (i.e., where causality cannot be inferred directly from the raw data), as in this case study, analyzing historical data.

In summary, the use of latent variable-based methods allowed the efficient use of the Six Sigma methodology in a batch production process where this could not have been done using a traditional Six Sigma toolkit, which lead to significant short- and long-term savings, in addition to the implementation of a more robust monitoring system.

5. Conclusions

Traditional Six Sigma statistical toolkit, mainly focused on classical statistical techniques (such as scatterplots, correlation coefficients, and linear regression models from experimental designs), is seriously handicapped for problem solving using process data coming from Industry 4.0. In this context, abundant historical process data involving hundreds/thousands of variables highly correlated with missing values are registered from daily production.

PCA can be used in this context as an exploratory tool not only to reduce the dimension of the original space and visualize the complex variables relationship but also to deal with missing data, identify patterns on data, trends, clusters, and outliers.

As data do not come from a DOE, input-output correlation does not mean necessarily causation, and classical predictive models (such as MLR and ML), proven to be very powerful in passive applications (i.e., predictions, process monitoring, fault detection, and diagnosis), cannot be used for extracting interpretable or causal models from historical data for process understanding, trouble-shooting, and optimization (active use), key goals of any Six Sigma project. This is the essence of the Box et al. (2005) warning [30]: predictive models based on correlated inputs must not be used for process optimization if they are built from observational data (i.e., data not coming from a DOE).

In contrast to classical MLR or ML techniques, PLS regression provides unique and causal models in the latent space even if data come from daily production process. These properties make PLS suitable for process optimization no matter where the data come from.

Therefore, Six Sigma's DMAIC methodology can achieve competitive advantages, efficient decision-making and problem-solving capabilities within the Industry 4.0 context by incorporating latent variable-based techniques, such as principal component analysis and partial least squares regression, into the statistical toolkit leading to Multivariate Six Sigma.

Author Contributions: Conceptualization, A.F.; methodology, A.F., J.B.-F. and D.P.-L.; software, J.B.-F. and D.P.-L.; validation, J.B.-F. and L.T.d.S.d.O.; formal analysis, J.B.-F., L.T.d.S.d.O. and D.P.-L.; data curation, L.T.d.S.d.O. and D.P.-L.; writing—original draft preparation, J.B.-F. and D.P.-L.; writing—review and editing, A.F.; visualization, J.B.-F., L.T.d.S.d.O. and D.P.-L.; supervision, A.F.; project administration, A.F.; funding acquisition, A.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external public funding.

Acknowledgments: The authors would like to acknowledge the invaluable help received from José Antonio Rojas, Antonio José Ruiz, Tim Ridgway, Ignacio Martin, and Steve Dickens, as well as the technicians at the plant, who helped greatly in the acquisition of the data, investigation at the plant, and essential understanding of the process. Without their knowledge and experience, the success of this project would have not been possible.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Linderman, K.; Schroeder, R.G.; Zaheer, S.; Choo, A.S. Six Sigma: A goal-theoretic perspective. *J. Oper. Manag.* **2003**, *21*, 193–203. [CrossRef]
2. Grima, P.; Marco-Almagro, L.; Santiago, S.; Tort-Martorell, X. Six Sigma: Hints from practice to overcome difficulties. *Total Qual. Manag. Bus. Excell.* **2014**, *25*, 198–208. [CrossRef]
3. Reis, M.S.; Gins, G. Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis. *Processes* **2017**, *5*, 35. [CrossRef]
4. Ferrer, A. Multivariate Statistical Process Control Based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly Process. *Qual. Eng.* **2007**, *19*, 311–325. [CrossRef]
5. Peruchi, R.S.; Rotela Junior, P.; Brito, T.G.; Paiva, A.P.; Balestrassi, P.P.; Mendes Araujo, L.M. Integrating Multivariate Statistical Analysis Into Six Sigma DMAIC Projects: A Case Study on AISI 52100 Hardened Steel Turning. *IEEE Access* **2020**, *8*, 34246–34255. [CrossRef]
6. Ismail, A.; Mohamed, S.B.; Juahir, H.; Toriman, M.E.; Kassim, A. DMAIC Six Sigma Methodology in Petroleum Hydrocarbon Oil Classification. *Int. J. Eng. Technol.* **2018**, *7*, 98–106. [CrossRef]
7. Jaeckle, C.M.; Macgregor, J.F. Product Design through Multivariate Statistical Analysis of Process Data. *AIChE J.* **1998**, *44*, 1105–1118. [CrossRef]
8. Höskuldsson, A. PLS regression methods. *J. Chemom.* **1988**, *2*, 581–591. [CrossRef]
9. Wold, S.; Sjostrom, M.; Eriksson, L. PLS-Regression—A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
10. MacGregor, J.F. Empirical Models for Analyzing "BIG" Data—What's the Difference? In Proceedings of the 2018 Spring Meeting and 14th Global Congress on Process Safety, Orlando, FL, USA, 22–26 April 2018; AIChE: New York, NY, USA, 2018.
11. García Muñoz, S.; MacGregor, J.F. Big Data: Success Stories in the Process Industries. *Chem. Eng. Prog.* **2016**, *112*, 36–40.
12. De Mast, J.; Lokkerbol, J. An analysis of the Six Sigma DMAIC method from the perspective of problem solving. *Int. J. Prod. Econ.* **2012**, *139*, 604–614. [CrossRef]
13. Tomba, E.; Facco, P.; Bezzo, F.; Barolo, M. Latent variable modeling to assist the implementation of Quality-by-Design paradigms in pharmaceutical development and manufacturing: A review. *Int. J. Pharm.* **2013**, *457*, 283–297. [CrossRef] [PubMed]
14. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
15. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
16. Bro, R.; Smilde, A.K. Principal component analysis. *Anal. Methods* **2014**, *6*, 2812–2831. [CrossRef]
17. Wold, H. Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*; Krishnaiah, P.R., Ed.; Academic Press: New York, NY, USA, 1966; pp. 391–420.
18. Tomba, E.; Barolo, M.; García-Muñoz, S. General Framework for Latent Variable Model Inversion for the Design and Manufacturing of New Products. *Ind. Eng. Chem. Res.* **2012**, *51*, 12886–12900. [CrossRef]
19. Kourti, T.; Macgregor, J.F. Multivariate SPC Methods for Process and Product Monitoring. *J. Qual. Technol.* **1996**, *28*, 409–428. [CrossRef]
20. Geladi, P.; Kowalski, B.R. Partial least-squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17. [CrossRef]
21. Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173. [CrossRef]

22. Wold, S.; Kettaneh-Wold, N.; MacGregor, J.F.; Dunn, K.G. Batch Process Modeling and MSPC. *Compr. Chemom.* **2009**, *2*, 163–197.
23. MacGregor, J.F.; Nomikos, P. Multivariate SPC Charts for Batch Monitoring Processes. *Technometrics* **1995**, *37*, 41–59.
24. Wold, S.; Kettaneh, N.; Fridén, H.; Holmberg, A. Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 331–340. [CrossRef]
25. Kourti, T. Abnormal situation detection, three-way data and projection methods; robust data archiving and modeling for industrial applications. *Annu. Rev. Control* **2003**, *27 II*, 131–139. [CrossRef]
26. González-Martínez, J.M.; De Noord, O.E.; Ferrer, A. Multisynchro: A novel approach for batch synchronization in scenarios of multiple asynchronisms. *J. Chemom.* **2014**, *28*, 462–475.
27. Kassidas, A.; MacGregor, J.F.; Taylor, P.A. Synchronization of Batch Trajectories Using Dynamic Time Warping. *AIChE J.* **1998**, *44*, 864–875. [CrossRef]
28. Camacho, J.; Pérez-Villegas, A.; Rodríguez-Gómez, R.A.; Jiménez-Mañas, E. Multivariate Exploratory Data Analysis (MEDA) Toolbox for Matlab. *Chemom. Intell. Lab. Syst.* **2015**, *143*, 49–57. [CrossRef]
29. González-Martínez, J.M.; Camacho, J.; Ferrer, A. MVBatch: A matlab toolbox for batch process modeling and monitoring. *Chemom. Intell. Lab. Syst.* **2018**, *183*, 122–133. [CrossRef]
30. Box, G.E.P.; Hunter, W.G.; Hunter, J.S. *Statistics for Experimenters: Design, Discovery and Innovation*, 2nd ed.; John Wiley and Sons: Hoboken, NJ, USA, 2005.